Further explanation of the effectiveness of voting methods: the game between margins and weights

Vladimir Koltchinskii[1], Dmitriy Panchenko[1], and Fernando Lozano[2]

[1] Department of Mathematics and Statistics, The University of New Mexico,
Albuquerque, NM, 87131, USA
{vlad, panchenk}@math.unm.edu
[2] Departamento de Ingeniería Electrónica, Universidad Javeriana,
Cr. 7 40-62, Bogotá, Colombia
fernando.lozano@javeriana.edu.co

Abstract. In this paper we present new bounds on the generalization error of a classifier f constructed as a convex combination of base classifiers from the class $\mathcal{H}$. The algorithms of combining simple classifiers into a complex one, such as boosting and bagging, have attracted a lot of attention. We obtain new sharper bounds on the generalization error of combined classifiers that take into account both the empirical distribution of "classification margins" and the "approximate dimension" of the classifier, which is defined in terms of weights assigned to base classifiers by a voting algorithm. We study the performance of these bounds in several experiments with learning algorithms.

1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample of n labeled training examples defined on a probability space $(\Omega, \Sigma, \mathbb{P})$. We assume that the examples are independent identically distributed copies of a random couple (X, Y), X being an "instance" in a measurable space $(S, \mathcal{A})$ and Y being a "label" taking values in $\{-1, 1\}$. Let P denote the distribution of the couple (X, Y). Given a measurable function f from S into $\mathbb{R}$, we use $\mathrm{sign}(f(x))$ as a predictor of the unknown label of an instance $x \in S$. We will call f a classifier of the examples from S. It is obvious that according to the above definition f predicts a label of x correctly if and only if $yf(x) > 0$. The quantity $m(x, y) = yf(x)$ will be called *the classification margin.* The probability $\mathbb{P}\{Yf(X) \le 0\} = P\{(x, y) : yf(x) \le 0\}$ defines *the generalization error* of the classifier f. The goal of learning (classification) is, given a set of training examples, to find a classifier f with a small generalization error.

Given a class of base (simple) classifiers $\mathcal{H}$, all voting algorithms produce a complex classifier that is a convex combination of base classifiers, i.e. belongs to the symmetric convex hull of $\mathcal{H}$:

$$\mathcal{F} := \mathrm{conv}(\mathcal{H}) := \Big\{ \sum_{i=1}^{N} \lambda_i h_i : N \ge 1, \lambda_i \in \mathbb{R}, \sum_{i=1}^{N} |\lambda_i| \le 1,\ h_i \in \mathcal{H} \Big\}.$$

The explanation of the effectiveness of voting algorithms requires the construction of a good bound on the generalization error $P\{yf(x) \leq 0\}$ uniformly over $\mathcal{F}$. The results we present here follow the line of research that started with the paper of Schapire, Freund, Bartlett and Lee [13]. The main difficulty one encounters trying to prove a uniform bound on the generalization error of a classifier from the convex hull $\mathrm{conv}(\mathcal{H})$ is that even if the original class $\mathcal{H}$ had a finite VC-dimension, the convex hull can be significantly more complex. In this case the standard techniques of the VC-theory can not be used directly. Recall that the main idea of this approach is based on the following easy bound

$$P\{(x,y) : yf(x) \leq 0\} \leq P_n\{(x,y) : yf(x) \leq 0\} + \sup_{C \in \mathcal{C}} [P(C) - P_n(C)],$$

where P_n is the empirical distribution of the training examples, i.e. for any set $C \subset S \times \{-1, 1\}$, $P_n(C)$ is the frequency of the training examples in the set C, $\mathcal{C} := \Big\{\{(x,y) : yf(x) \leq 0\} : f \in \mathcal{F}\Big\}$, and on further bounding the uniform (over the class $\mathcal{C}$) deviation of the empirical distribution P_n from the true distribution P. The methods that are used to construct such uniform deviation bounds belong to the theory of empirical processes and the crucial role is played by the VC-dimension, or by more sophisticated entropy characteristics of the class $\mathcal{C}$. For instance, if $m^{\mathcal{C}}(n)$ denotes the maximal number of subsets produced by intersecting a sample of size n with the class $\mathcal{C}$ (the so called shattering number), then the following bound holds (see [5], Theorem 12.6) for all $\varepsilon > 0$

$$\mathbb{P}\Big\{\exists f \in \mathcal{F} : P\{yf(x) \leq 0\} \geq P_n\{yf(x) \leq 0\} + \varepsilon\Big\} \leq 8 m^{\mathcal{C}}(n) e^{-n\varepsilon^2/32}.$$

It follows from this bound that the training error measures the generalization error of a classifier $f \in \mathcal{F}$ with the accuracy $O\Big(\sqrt{\frac{V(\mathcal{C}) \log n}{n}}\Big)$, where $V(\mathcal{C})$ is the VC-dimension of the class $\mathcal{C}$. In the case of classifiers with zero training error, the accuracy can be improved to $O\Big(\frac{V(\mathcal{C}) \log n}{n}\Big)$. The above bounds, however, do not apply directly to the case of the class $\mathcal{F} = \mathrm{conv}(\mathcal{H})$, which is of interest in applications to bounding the generalization error of the voting methods, since in this case typically $V(\mathcal{C}) = +\infty$. Even when one deals with a finite number of base classifiers in a convex combination (which is the case, say, with boosting after finite number of rounds), the VC-dimensions of the classes involved are becoming rather large, so the above bounds do not explain the generalization ability of boosting and other voting methods observed in numerous experiments.

In [13] Schapire et al. (see also [1],[2]) developed a new class of bounds on generalization error of a convex combination of classifiers, expressed in terms of empirical distribution of margins. They showed that for a given $\alpha \in (0,1)$ with probability at least $1 - \alpha$ for all $f \in \mathrm{conv}(\mathcal{H})$

$$P\{yf(x) \leq 0\} \leq \inf_{\delta} \Big[P_n\{yf(x) \leq \delta\} + \frac{C}{\sqrt{n}} \Big(\frac{V}{\delta^2} \log^2 \frac{n}{V} + \log \frac{1}{\alpha}\Big)^{1/2}\Big], \quad (1)$$

where V is the VC-dimension of $\mathcal{H}$. Choosing in the above bound the value of $\delta = \hat{\delta}(f)$ that solves the equation

$$\delta P_n\{yf(x) \le \delta\} = (V/n)^{1/2}$$

(which is nearly an optimal choice), one gets (ignoring the logarithmic factors) the generalization error of a classifier f from the convex hull of the order $O\Big(\frac{1}{\hat{\delta}(f)}\sqrt{\frac{V}{n}}\Big)$. Schapire et al. showed that in many experiments voting methods tended to classify the majority of examples not only correctly but with a high confidence, i.e. with a large margin, which means that one can expect $\hat{\delta}$ to be reasonably large and, therefore, the bound becomes meaningful.

In [8],[9] using the methods of theory of Empirical, Gaussian and Rademacher Processes (concentration inequalities, symmetrization, comparison inequalities) we generalized and refined this type of bounds. In our first result we do not immediately assume that $\mathcal{H}$ is a VC-class but propose to measure the complexity of the class in terms of what we call the Rademacher complexity function

$$R_n(\mathcal{H}) := \mathbb{E}\sup_{h\in\mathcal{H}} |n^{-1}\sum_{j=1}^n \varepsilon_j h(X_j)|,$$

where ε_j, $j = 1, \dots, n$ are i.i.d. Rademacher random variables. Then similarly to (1) we prove that for all $\alpha \in (0,1)$, with probability at least $1-\alpha$, $\forall f \in \mathrm{conv}(\mathcal{H})$

$$P\{yf(x) \le 0\} \le \inf_{\delta\in(0,1]} \Big[P_n(yf(x) \le \delta) + \frac{8}{\delta}R_n(\mathcal{H}) + \frac{1}{\sqrt{n}}\Big(\log\log_2 \frac{2}{\delta}\Big)^{1/2} + \frac{1}{\sqrt{n}}\Big(\frac{1}{2}\log\frac{2}{\alpha}\Big)^{1/2}\Big]. \tag{2}$$

The theory of empirical processes provides a number of bounds for $R_n(\mathcal{H})$ in terms of different characteristics of complexity of the class $\mathcal{H}$. For example, in the case when $\mathcal{H}$ is a VC-class with VC-dimension V one has the following bound ([16]) $R_n(\mathcal{H}) \le C\Big(\frac{V}{n}\Big)^{1/2}$ which shows that (2) improves (1).

Next, we suggested a way to improve these bounds even further. Again we save the case $\mathrm{conv}(\mathcal{H})$ where $\mathcal{H}$ is a VC-class as an example and work with the general assumption on the growth of random entropies of a class $\mathcal{F}$ to which the classifier belongs. Given a metric space (T, d), we denote $H_d(T; \varepsilon)$ the ε-entropy of T with respect to d, i.e. $H_d(T;\varepsilon) := \log N_d(T;\varepsilon)$, where $N_d(T;\varepsilon)$ is the minimal number of balls of radius ε covering T. If Q is a probability measure on $(S;\mathcal{A})$, $d_{Q,2}$ will denote the metric of the space $L_2(S;dQ)$: $d_{Q,2}(f;g) := (Q|f-g|^2)^{1/2}$. We assume that for some $\alpha \in (0,2)$

$$H_{d_{P_n,2}}(\mathcal{F};u) \le D_n^2 u^{-\alpha}, \;\; u > 0 \quad \text{a.s.}, \tag{3}$$

where $D_n = D_n(X_1, \dots, X_n)$ is a function of training examples such that $\mathbb{E}D_n < \infty$. If $\mathcal{F} = \mathrm{conv}(\mathcal{H})$ and $\mathcal{H}$ is a VC-class with VC-dimension V, then (3) holds with $\alpha = 2(V-1)/V$ and D_n is a constant that depends on the VC-dimension only (see [16]). To formulate one of the results that was obtained in [8] we need the following definitions. Given $\gamma > 0$, we define a γ-margin and an empirical γ-margin of the function f by

$$\delta_n(\gamma; f) := \sup\Big\{\delta \in (0,1) : \delta^\gamma P\{yf(x) \le \delta\} \le n^{-1+\frac{\gamma}{2}}\Big\},$$

$$\hat{\delta}_n(\gamma; f) := \sup\Big\{\delta \in (0,1) : \delta^\gamma P_n\{yf(x) \le \delta\} \le n^{-1+\frac{\gamma}{2}}\Big\}.$$

We proved in [8] that under the condition (3) for any $\gamma \ge 2\alpha/(2+\alpha)$ there exist constants $A, B > 0$ such that for n large enough

$$\forall f \in \mathcal{F} : A^{-1}\hat{\delta}_n(\gamma; f) \le \delta_n(\gamma; f) \le A\hat{\delta}_n(\gamma; f).$$

with probability at least

$$1 - B\log_2\log_2 n \exp\{-n^{\frac{\gamma}{2}}/2\}.$$

What this result says is that with high probability the "true" γ-margin and the empirical γ-margin are within a constant factor of each other. One can notice that the definition of the γ-margins contains the bound on the generalization error of a function f. It easily follows that with high probability for some constant A' and for an arbitrary $f \in \mathcal{F}$

$$P\{yf(x) \le 0\} \le \frac{A'}{n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma}. \tag{4}$$

It's easy to check that the quantity $(n^{1-\gamma/2}\hat{\delta}_n(\gamma; f)^\gamma)^{-1}$ is *decreasing* as γ decreases from 1 to 0. The previous bounds (1) and (2) corresponded to the worst case $\gamma = 1$. As we already mentioned above, if $\mathcal{F} = \mathrm{conv}(\mathcal{H})$ and $\mathcal{H}$ is a VC-class with VC-dimension V (this includes all voting methods), then $\alpha = 2(V-1)/V < 2$ and $\gamma = 2(V-1)/(2V-1) < 1$, improving the previous bounds.

Though qualitatively the γ-bounds constitute an improvement upon (1) and (2), when the VC-dimension V is large γ can be very close to 1. Our experiments [10] showed that, in the case of the classifiers obtained in consecutive rounds of boosting, the bounds on the generalization error in terms of γ-margins hold even for much smaller values of γ. This allows one to conjecture that such classifiers belong, in fact, to a class $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ whose entropy might be much smaller than the entropy of the whole convex hull. The problem, though, is that it is practically impossible to identify such a class prior to experiments, leaving the question of how to choose the values of γ for which the bounds hold open.

2 Balancing approximate dimensions and margins

In an attempt to capture a smaller subclass of the convex hull to which the classifier belongs we develop a new approach. Namely, we suggest an adaptive bound on the generalization error of a classifier produced by a specific procedure that in some sense tries to localize the location of the classifier inside the convex hull. We consider an unnested family of subsets of $\mathrm{conv}(\mathcal{H})$ that are defined in terms of weights (decay of weights) of the convex combination; the conditions on weights imply the bounds on the random entropy of these subclasses which in turn are used to prove the bounds on the generalization error. For example, the subset which corresponds to convex combinations with very few large weights must have a smaller complexity than the whole convex hull and, therefore, enjoy a sharper bound. The classifier represented by a convex combination may belong to many of these subsets which leads to a family of bounds. The adaptive bound suggested below is based on "optimizing" the bounds over the whole family.

Now we will give precise definitions and make these ideas rigorous. Let $\mathcal{H}$ be a VC-class of measurable functions from $(S, \mathcal{A})$ into $\{-1, 1\}$ with VC-dimension V. Let $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$. For a function $f \in \mathcal{F}$ and a number $\Delta \in [0, 1]$, we define *the approximate Δ-dimension* of f as the smallest integer number $d \geq 0$ such that there exist $N \geq 1$, functions $h_j \in \mathcal{H}$, $j = 1, \ldots, N$ and numbers $\lambda_j \in \mathbb{R}$, $j = 1, \ldots, N$ satisfying the conditions $f = \sum_{j=1}^{N} \lambda_j h_j$, $\sum_{j=1}^{N} |\lambda_j| \leq 1$ and $\sum_{j=d+1}^{N} |\lambda_j| \leq \Delta$. The Δ-dimension of f will be denoted by $d(f; \Delta)$.

Let $\alpha := 2(V-1)/V$ and $\Delta_f = \{\Delta \in [0, 1] : d(f; \Delta) \leq n\}$. Define

$$\varepsilon_n(f;\delta) := \inf_{\Delta\in\Delta_f}\Big[\frac{d(f;\Delta)}{n}\log\frac{ne^2}{\delta d(f;\Delta)} + \Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}\Big]\bigvee\frac{2\log n}{n}, \qquad (5)$$

where $a \vee b := \max\{a, b\}$. Let

$$\hat{\delta}_n(f) := \sup\Big\{\delta \in (0, 1/2) : P_n\{yf(x) \leq \delta\} \leq \varepsilon_n(f;\delta)\Big\}.$$

Theorem 1. *There exist constants $A, B > 0$ such that for all $0 < t < n^{\frac{\alpha}{2+\alpha}}$*

$$\forall f \in \mathcal{F} \;\; P\{yf(x) \leq \frac{\hat{\delta}_n(f)}{4}\} \leq \; A\Big(\varepsilon_n(f; \frac{\hat{\delta}_n(f)}{2}) + \frac{t}{n}\Big)$$

with probability at least $1 - Be^{-t/4}$.

To understand this bound let us look at the definition of $\varepsilon_n(f;\delta)$. First of all, if instead of minimizing over Δ one sets $\Delta = 1$, then, since $d(f, 1) = 0$, the bound becomes equivalent to the previous γ-bound (4), which means that the bound of the theorem improves the γ-bound. For a fixed Δ, the two terms in the definition of $\varepsilon_n(f;\delta)$ correspond to two parts of the combined classifier. The first term corresponds to the sum of $d(f, \Delta)$ base classifiers with the largest weights and the form of the bound basically coincides with the standard VC-dimension based bound in the zero error case. The second term corresponds to an "improper" convex combination of classifers with the smallest weights (the number of them is not limited), and the form of the bound is determined by the complexity of the whole convex hull, only scaled by a factor of Δ. It is clear that if a voting algorithm produces a convex combination in which there are very few classifiers with large weights, then the bound of the theorem can improve upon (4) significantly. Another way to say it is that the faster is the weight decay in the convex combination, the smaller is the complexity of the corresponding subset of the convex hull and the sharper is the bound.

As an example, we can assume that the algorithm produces a classifier with polynomial or exponential decay of the weights, which allows us to minimize (5) explicitly over Δ to see how the bound looks like. If $\mathcal{F} \subset \mathrm{conv}(\mathcal{H})$ is a class of functions such that for some $\beta > 0$

$$\sup_{f\in\mathcal{F}} d(f;\Delta) = O(\Delta^{-\beta}), \tag{6}$$

then with "high probability" for any classifier $f \in \mathcal{F}$ the upper bound on its generalization error becomes of the order

$$\frac{1}{n^{1-\gamma\beta/2(\gamma+\beta)}\hat{\delta}_n(f)^{\gamma\beta/(\gamma+\beta)}},$$

(which, of course, improves a more general bound in terms of γ-margins; the general bound corresponds to the case $\beta = +\infty$). The condition (6) means that the weights of the convex combination decrease polynomially fast, namely, $|\lambda_j| = O(j^{-\alpha})$, $\alpha = 1 + \beta^{-1}$. The case of exponential decrease of the weights is described by the condition

$$\sup_{f\in\mathcal{F}} d(f;\Delta) = O(\log\frac{1}{\Delta}). \tag{7}$$

In this case the upper bound becomes of the order $\frac{1}{n}\log^2\frac{n}{\hat{\delta}_n(f)}$.

The complete proofs of the results require many more pages than it is available for us here. They can be found in our papers [8], [9] that are available online at www.boosting.org. Here we only give the sketch of the proof of Theorem 1, which is the main result of this paper. In the first and main part of the proof we consider a family of classes $\mathcal{F}_{d,\Delta}$ parametrized by two parameters d and Δ, and prove a uniform bound on the generalization error over any fixed class in this family. In the second part of the proof we make this bound adaptive, which means that if a function f belongs to more than one class in the family, then one can choose a class that provides the best bound.

Sketch of proof. Let us fix $\delta \in (0, 1/2]$. For any function f we denote $d(f) := d(f, \bar{\Delta})$, where $\bar{\Delta}$ is such that the infimum in the definition (5) is attained at $\bar{\Delta}$. For a fixed δ we consider a partition of $\mathcal{F}$ into two classes $\mathcal{F}_1^\delta$ and $\mathcal{F}_2^\delta = \mathcal{F} \setminus \mathcal{F}_1^\delta$, where $\mathcal{F}_1^\delta := \{f : d(f) = 0\}$ (note that $d(f)$ depends on δ). The fact that $f \in \mathcal{F}_1^\delta$ means that the weights of the classifier f are distributed "uniformly" and in this case the bound of Theorem 1 does not improve (4). The family of classes that we use to localize the classifier f is defined as follows:

$$\mathcal{F}_{d,\Delta} := \{f \in \mathcal{F}_2^\delta : d(f;\Delta) \le d\}.$$

If $f \in \mathcal{F}_{d,\Delta}$ and Δ is small then it means that the "voting power" is concentrated in the faction consisting of the first d base classifiers of the convex combination. First of all we estimate the complexity of this class. The definition of $\mathcal{F}_{d,\Delta}$ implies the following bound on the random entropy:

$$H_{d_{P_n,2}}(\mathcal{F}_{d,\Delta}; u) \le K\Big[d \log\frac{e}{u} + \Big(\frac{\Delta}{u}\Big)^{\alpha}\Big] \text{ for } u \le 1, \tag{8}$$

where K is a positive constant. To prove it one has to represent $\mathcal{F}_{d,\Delta}$ as

$$\mathcal{F}_{d,\Delta} \subseteq \mathcal{H}^d + \Delta \mathrm{conv}\mathcal{H}.$$

Next we use the complexity estimate (8) to prove a uniform bound on the generalization error of classifiers in $\mathcal{F}_{d,\Delta}$. The proof is based on an iterative application of Talagrand's concentration inequality for empirical processes (see [15], [12]) which allows us to measure the size of $\mathcal{F}_{d,\Delta}$ correctly and make a better inference about the generalization error of elements of this class. Let us first formulate the bound precisely. Let $1 \le d \le n$ and denote

$$\varepsilon_n(d;\delta;\Delta) := \Big[\frac{d}{n}\log\frac{ne^2}{\delta d} + \Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}\Big] \bigvee \frac{2\log n}{n}.$$

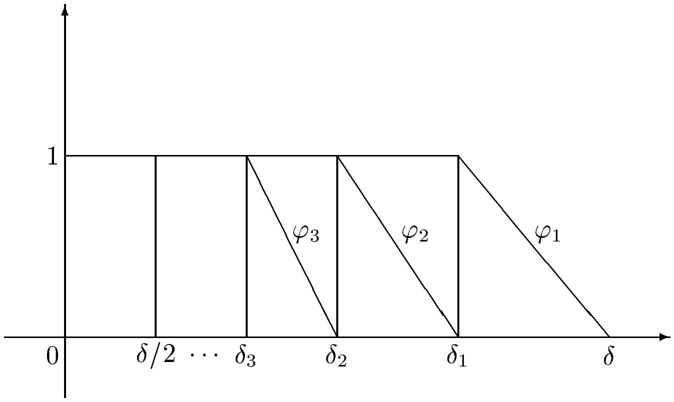
We prove that there exist constants $A, B > 0$ such that the following event

$$\forall f \in \mathcal{F}_{d,\Delta} \,:\, P_n\{m \le \delta\} \le \varepsilon_n(d;\delta;\Delta) \;\text{ implies }\; P\{m \le \frac{\delta}{2}\} \le A\varepsilon_n(d;\delta;\Delta), \tag{9}$$

where $m(x,y) = yf(x)$, occurs with probability at least

$$1 - B\Big(\frac{\delta d}{n}\Big)^{d/4} \exp\Big\{-\frac{1}{4}\Big(\sqrt{n}\frac{\Delta}{\delta}\Big)^{2\alpha/(\alpha+2)}\Big\}.$$

To proceed from here, one has to carefully eliminate the dependence on d, Δ and δ, and as a result to get the adaptive bound of Theorem 1. This constitutes the second part of the proof.



Let us now describe the iterative localization scheme that we used to prove (9). For a fixed δ, we choose (in a rather special way) a finite decreasing sequence $\delta_j, 1 \leq j \leq N$ such that $\delta_j < \delta$ and $\delta_N = \delta/2$. We consider the functions $\varphi_j, j \geq 1$ which are defined as shown in the figure above and which play the role of continuous approximations of indicator step functions $I(x \leq \delta_j)$. For simplicity of notations, we will suppress the variable Y in the couple (X, Y) and assume that a function f denotes it's margin, instead of writing $yf(x)$. We start by assuming that the empirical distribution of the margin at the point δ is small, i.e. $P_n(f \leq \delta) \ll 1$, and after N iterations arrive at the bound of the same magnitude for $P(f \leq \delta/2)$, improving it at each step. Let us show how the iterative step is implemented. Skipping the first step we assume that we already showed that with high probability $P(f \leq \delta_1) \leq r_1 \ll 1$ (it can be done similarly to what we do below, just set $\sigma_1 := 1$). This means that f belongs to a (small) subset of $\mathcal{F}$, namely, $\mathcal{F}_1 = \{f \in \mathcal{F} : P(f \leq \delta_1) \leq r_1\}$. The following series of inequalities is clear (we use the notations $\|Z\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |Z(f)|$, $Z : \mathcal{F} \mapsto \mathbb{R}$ and $\varphi(\mathcal{F}) := \{\varphi(f) : f \in \mathcal{F}\}$):

$$P(f \leq \delta_2) \leq P\varphi_2(f) \leq P_n\varphi_2(f) + \|P_n - P\|_{\varphi_2(\mathcal{F}_1)} \leq P_n(f \leq \delta_1) + \|P_n - P\|_{\varphi_2(\mathcal{F}_1)}.$$

Now Talagrand's concentration inequality for empirical processes implies that for a fixed ε with high probability (at least $1 - e^{-n\varepsilon/2}$)

$$\|P_n - P\|_{\varphi_2(\mathcal{F}_1)} \leq K(\mathbb{E}\|P_n - P\|_{\varphi_2(\mathcal{F}_1)} + \sqrt{\sigma_1^2 \varepsilon} + \varepsilon)$$

for some constant $K > 0$ and $\sigma_1^2 = \sup_{\mathcal{F}_1} P\varphi_2^2(f) \leq P(f \leq \delta_1) \leq r_1 \ll 1$ according to the bound from the previous step. Dudley's entropy bound (see [6]) implies in this case that

$$\mathbb{E}\|P_n - P\|_{\varphi_2(\mathcal{F}_1)} \leq K\mathbb{E}\int_0^{\sigma_1} H^{1/2}_{d_{P_n,2}}(\varphi_2(\mathcal{F}_1), u)du,$$

which is further bounded using the entropy condition (8). Collecting all these estimates one can show that $P(f \leq \delta_2) \leq r_2 \leq r_1$. Therefore, with high probability $f \in \mathcal{F}_2 = \{f \in \mathcal{F} : P(f \leq \delta_2) \leq r_2\}$, which in terms of generalization error is "smaller" than $\mathcal{F}_1$. As one can see this improvement was possible because of the fact that Talagrand's inequality measures the size of the class via σ_1. Now the similar argument can be iterated until the bound reaches (up to a multiplicative constant) the optimal fixed point of this recursive procedure, which is precisely formulated in (9).

3 Experiments

In this section we present the results of several experiments we conducted to test the ability of our bounds to predict the classification error of combined classifiers. Even though all the steps of the proofs of our results allow one to use explicit constants, the values of the constants will be too large due to the generality of the methods of empirical processes upon which our proof is heavily based. For example, the constants in Talagrand's concentration inequality are known (see [11]), but they are most likely far from being optimal. Therefore, we will simply use the quantities $(n^{1-\gamma/2}\hat{\delta}_n(\gamma;f)^{\gamma})^{-1}$ and $\varepsilon_n(f;\hat{\delta}_n(f))$ instead of the upper bounds we actually proved, and we will refer to them as γ-bound and Δ-bound correspondingly.

We first describe the experiments with a "toy" problem which is simple enough to allow one to compute the generalization error exactly. Namely, we consider a one dimensional classification problem in which the space of instances S is an interval $[0,1]$ and, given a concept $C_0 \subset S$ which is a finite union of disjoint intervals, the label y is assigned to a point $x \in S$ according to the rule $y = f_0(x)$, where f_0 is equal to $+1$ on C_0 and to -1 on $S \setminus C_0$. We refer to this problem as the *intervals problem* (see also [7]). Note that for the class of decision stumps we have $V(\mathcal{H}) = 2$ (since $\mathcal{H} = \{I_{[0,b]} : b \in [0,1]\} \cup \{I_{[b,1]} : b \in [0,1]\}$), and according to the results above the values of γ in $[2/3, 1)$ provide valid bounds on the generalization error in terms of the γ-margins. In our experiments, the set C_0 was formed by 20 equally spaced intervals and the training set of size 1000 was generated by the uniform distribution on $[0,1]$. We ran Adaboost for 500 rounds and computed at each round the true generalization error of the combined classifier and the bounds for different values of γ.

In figure 1 we plot the true classification error and the γ-bounds for $\gamma = 1$, 0.8 and 2/3 against the number of iteration of Adaboost. The bound for $\gamma = 1$ corresponds to the previously known bounds (1) and (2) and as expected is inferior to the γ-bound for smaller values of γ. In figure 2 we compare the γ-bound for the best admissible $\gamma = 2/3$ with the Δ-bound of Theorem 1. As one can see when the number of iterations is small the Δ-bound takes advantage of the first "finite dimensional" term in the definition (5) since $d(f, \Delta)$ is small. When the number of iterations increases, the Δ-bound is gradually gravitating toward the more conservative γ-regime which means that the optimal value of Δ is increasing to 1. It seems unnatural that the bound increases while the true error decreases but as we will show later it can be simply a question of assigning different relative weights to two terms in the definition of the Δ-bound.

We also computed the bounds for more complex simulated data sets as well as for real data sets in which the same type of behavior was observed. We show the results for the Twonorm Data Set and the King Rook vs. King Pawn Data Set, using Adaboost and Bagging in figures 3-6. The Twonorm Data Set (taken from [4]) is a simulated 20 dimensional data set in which positive and negative training examples are drawn from the multivariate normal distributions with unit covariance matrix centered at $(2/\sqrt{20}, \ldots, 2/\sqrt{20})$ and $(-2/\sqrt{20}, \ldots, -2/\sqrt{20})$, respectively. The King Rook vs. King Pawn Data Set is a real data set from the UCI Irvine repository [3]. It is a 36 dimensional data set with 3196 samples.

As before, we used the decision stumps as base classifiers. An upper bound on $V(\mathcal{H})$ for the class $\mathcal{H}$ of decision stumps in R^d is given by the smallest n such that $2^{n-1} \geq (n-1)d + 1$. In each case we computed the Δ-bound and the γ-bounds for $\gamma = 1$ and for the smallest γ allowed by the theory ($\gamma_{\min}$). For

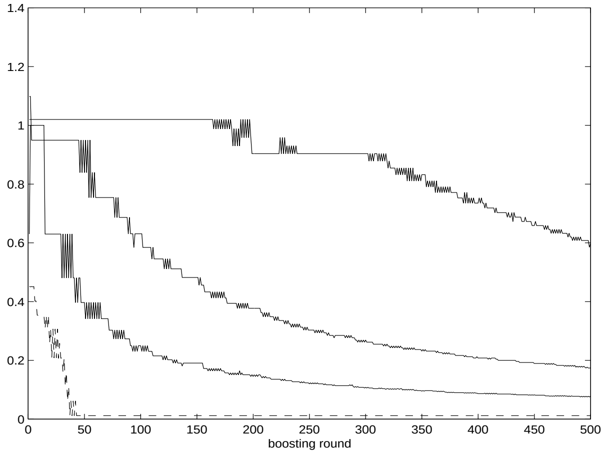


Fig. 1. Comparison of the generalization error (dashed line) with $(n^{1-\gamma/2}\hat{\delta}_n(\gamma;f)^\gamma)^{-1}$ for $\gamma = 1, 0.8$ and $2/3$ (solid lines, top to bottom)

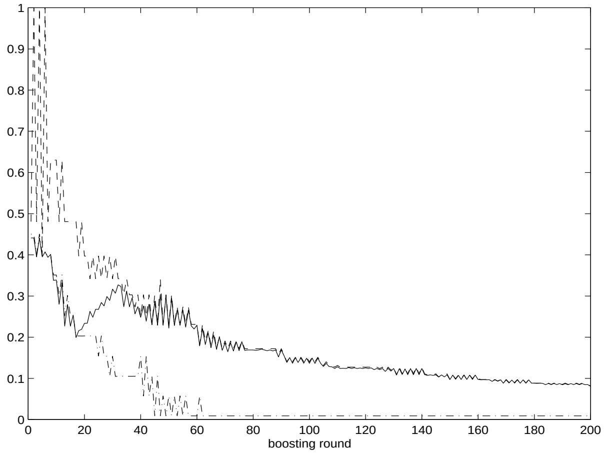


Fig. 2. Test error and bounds vs. number of classifiers for the intervals problem with a sample size of 1000. Test error (dot-dashed line), γ-margin bound with $\gamma = 2/3$ (dashed line), and Δ-bound (solid line).

the Twonorm Data Set, we estimated the generalization error by computing the empirical error on an indepedently generated set of 20000 observations. For the King Rook vs. King Pawn Data Set, we randomly selected 90% of the data for training and used the remaining 10% to compute the test error. The experiments were averaged over 10 repetitions. One can observe a similar two-regime behavior of the Δ-bound as in the intervals problem.

We also show that by slightly changing the definition of Δ-bound one can obtain in some cases a surprisingly accurate prediction of the shape of the generalization curve. The fact that both terms in the definition of $\varepsilon_n(f,\Delta)$ have weight 1 is related to lack of information about the values of the constants involved in the bounds. More subtle analysis can lead to a more general definition in which the weights of two terms might differ, for example

$$\varepsilon_{n,\zeta,K}(f;\delta) := K \inf_{\Delta\in\Delta_f}\Big[\zeta\frac{d(f;\Delta)}{n}\log\frac{ne^2}{\delta d(f;\Delta)} + (1-\zeta)\Big(\frac{\Delta}{\delta}\Big)^{\frac{2\alpha}{\alpha+2}} n^{-\frac{2}{\alpha+2}}\Big],$$

where $\zeta \in (0,1)$ and $K > 0$. Our goal in Theorem 1 was to understand the dependence of the bounds on the parameters of the problem and we were not concerned with the constants, but, in general, it would be more accurate to state the bound in this "weighted" form. In figures 7 and 8 we show the behavior of the modified Δ-bound for $\zeta = 0.1$ and 0.4. One can see that for a small value of ζ the "two-regime" behavior disappears and the bounds capture the shape of the true generalization curve (it should be emphasized that the value $\zeta = 0.1$ was determined based on an experiment with a toy example described above and was used for other data sets showing reasonable results in most of the experiments).

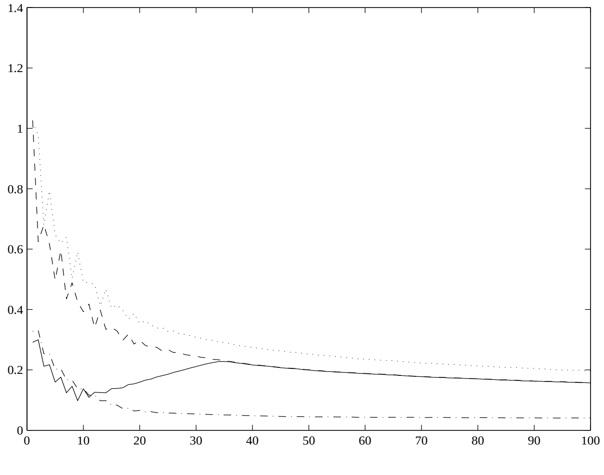


Fig. 3. Test error and bounds vs. number of classifiers for the twonorm data set using Adaboost. Test error (dot-dashed line), γ-margin bound with $\gamma = 1$ (dotted line), and $\gamma = \gamma_{\min}$ (dashed line), and Δ-bound (solid lines)

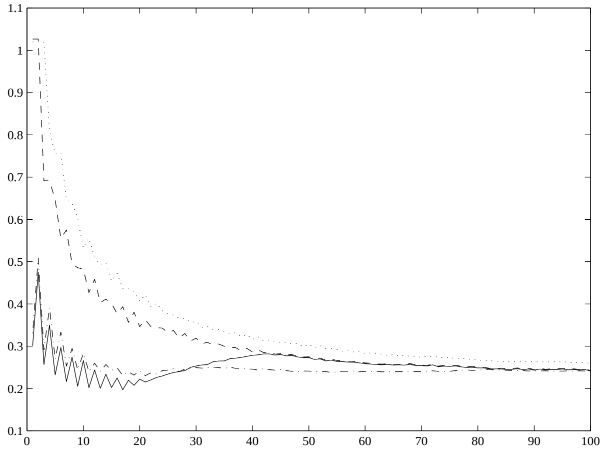


Fig. 4. Test error and bounds vs. number of classifiers for the twonorm data set using bagging. Test error (dot-dashed line), γ-margin bound with $\gamma = 1$ (dotted line), and $\gamma = \gamma_{\min}$ (dashed line), and Δ-bound (solid line)

4 Future goals

An obvious goal of future research is to identify and, if possible, to optimize the constants in the bounds we proved. Another goal is to develop a more subtle definition of approximate dimension of classifiers in the convex hull that takes into account the closeness of base classifiers in convex combinations (for instance, the closeness in the empirical distance $d_{P_n,2}$). This can further reduce the dimensionality of the classifiers and result in better bounds on generalization error of voting algorithms.

References

1. Anthony, M. and Bartlett, P. (1999) Neural Network Learning: Theoretical Foundations. Cambridge University Press.
2. Bartlett, P. (1998) The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network. *IEEE Transactions on Information Theory,* 44, 525–536.
3. Blake, C., Merz, C. (1998) UCI repository of machine learning databases. URL: http://www.ics.uci.edu/ mlearn/MLRepository.html.
4. Breiman, L. (1998) Arcing Classifiers. *The Annals of Statistics,* 26(3).
5. Devroye, L., Györfi, L. and Lugosi, L. (1996) A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.
6. Dudley, R.M. (1999) Uniform Central Limit Theorems. Cambridge University Press.

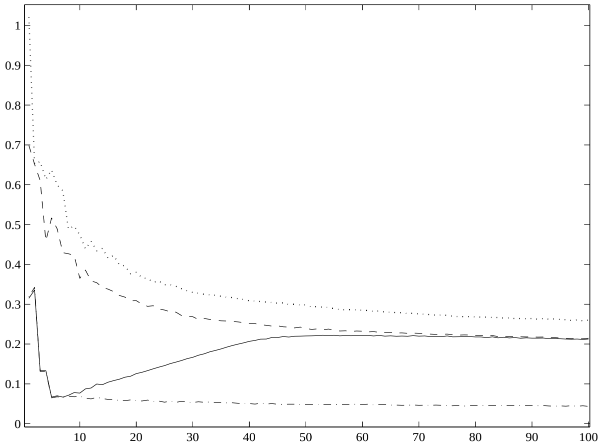


Fig. 5. Test error and bounds vs. number of classifiers for the King Rook vs. King Pawn data set using Adaboost. Test error (dot-dashed line), γ-margin bound with $\gamma = 1$ (dotted line), and $\gamma = \gamma_{\min}$ (dashed line), and Δ-bound (solid line)

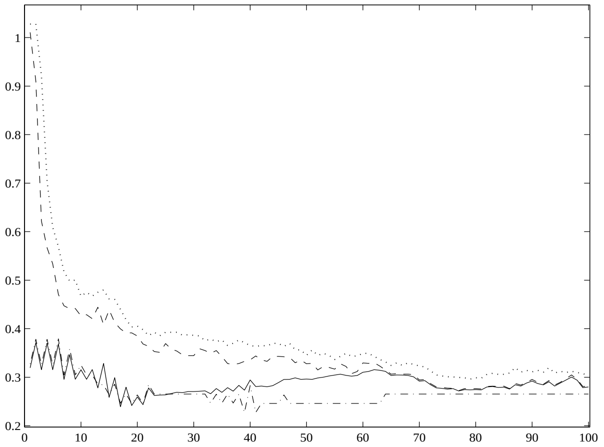


Fig. 6. Test error and bounds vs. number of classifiers for the King Rook vs. King Pawn data set using bagging Test error (dot-dashed lines), γ-margin bound with $\gamma = 1$ (dotted line), and $\gamma = \gamma_{\min}$ (dashed line), and Δ-bound (solid line)

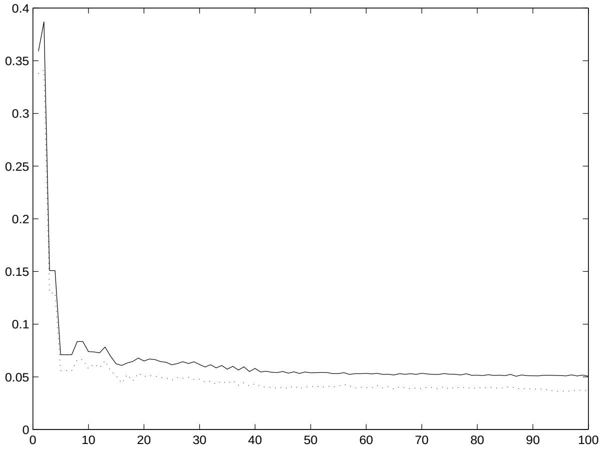


Fig. 7. Δ-Bound with $\zeta = 0.1$ (solid line), and test error (dotted line) for the King Rook vs. King Pawn data set

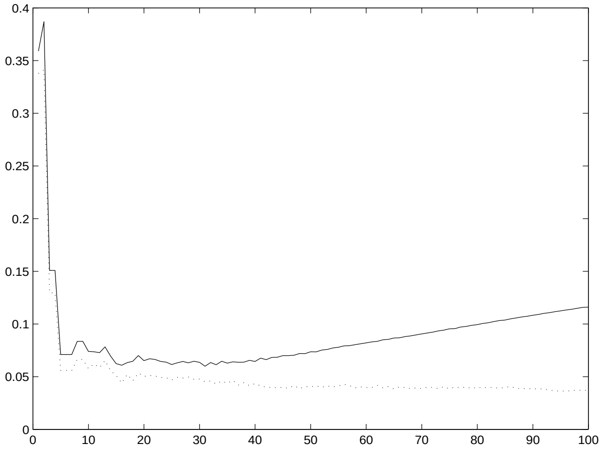


Fig. 8. Δ-Bound with $\zeta = 0.4$ (solid line), and test error (dotted line) for the King Rook vs. King Pawn data set

7. Kearns, M., Mansour, Y., Ng, A., Ron, D. (1997) An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning,* 27(1)
8. Koltchinskii, V. and Panchenko, D. (2000) Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. To appear in *Ann. Statist.*
9. Koltchinskii, V., Panchenko, D. and Lozano, F. (2000) Bounding the Generalization Error of Convex Combinations of Classifiers: Balancing the Dimensionality and the Margins. Preprint.
10. Koltchinskii, V., Panchenko, D. and Lozano, F. (2000) Some New Bounds on the Generalization Error of Combined Classifiers. *Advances in Neural Information Processing Systems 13: Proc. of NIPS'2000.*
11. Massart, P. (2000) About the Constants in Talagrand's Concentration Inequalities for Empirical Processes. *Ann. Probab.,* 28(2).
12. Panchenko, D. (2001) A Note on Talagrand's Concentration Inequality. To appear in *Electron. J. Probab.*
13. Schapire, R., Freund, Y., Bartlett, P. and Lee, W.S. (1998) Boosting the Margin: A New Explanation of Effectiveness of Voting Methods. *Ann. Statist.* 26, 1651–1687.
14. Talagrand, M. (1996a) A New Look at Independence. *Ann. Probab.,* 24, 1-34.
15. Talagrand, M. (1996b) New Concentration Inequalities in Product Spaces. *Invent. Math.,* 126, 505-563.
16. van der Vaart, A.W. and Wellner, J.A. (1996) Weak Convergence and Empirical Processes. With Applications to Statistics. Springer-Verlag, New York.